DOCUMENT RESUME

ED 246 107 TM 840 386

AUTHOR Reckase, Mark D.; McKinley, Robert L.
TITLE Item Difficulty Reconsidered: An IRT Perspective.
PUB DATE Apr 84
NOTE 13p.; Paper presented at the Annual Meeting of the American Educational Research Association (68th, New Orleans, LA, April 23-27, 1984).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Difficulty Level; Evaluation Methods; Item Analysis; *Latent Trait Theory; Test Construction; *Test Items

ABSTRACT

A new indicator of item difficulty, which identifies effectiveness ranges, overcomes the limitations of other item difficulty indexes in describing the difficulty of an item or a test as a whole and in aiding the selection of appropriate ability level items for a test. There are three common uses of the term "item difficulty": (1) the probability of a correct answer to a test item, based on the perception of the person being tested; (2) a group summary statistic that tells what proportion of a group can answer an item correctly; and (3) a measure of a location along a difficulty continuum based on item response theory (IRT) methodology. Using formulas based on IRT methodology, the new indicator measures the range of the score scale in which the item was effective and the point where the item is most informative. Unlike other methods, it is sample independent and can be used more directly for test construction by representing each item as a block, then selecting and stacking these blocks to create the distribution of test items on an ability scale. (ES)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED246107

Item Difficulty Reconsidered: An IRT Perspective

Mark D. Reckase and Robert L. McKinley
The American College Testing Program

The term "item difficulty" is commonly used in three different ways. First, from the point of view a person taking a test item, the item difficulty is related to the probability that he or she has of answering the test item correctly. If that probability is low, it is considered a difficult item. If the probability is high, the item is considered easy. This judgement of item difficulty may differ substantially from person to person. The statistical indication of the item difficulty for an individual is the probability of correct response. This index will also differ from person to person. The judgement of difficulty and the statistical index do not require any assumptions about the test item, except that there is a unique correct response.

The second use of the term "item difficulty" is as a group summary statistic that tells what proportion of a group can answer a test item correctly. This statistic can be thought of as the average of the probabilities of correct response for all of the members of a group. It will vary from group to group and it does not require any specific assumptions about the test item. However, knowing the group difficulty statistic, the traditional p index of difficulty, tells little about the difficulty of an item for a particular person. To determine an individual's probability of a correct response from the group p-value, the distribution of ability for the group and the position of the individual within the group must be known. Making the conversion from the group to individual difficulty statistic requires several assumptions about the ability being measured and its relationship to performance on the test item.

The third use of the term "item difficulty" is as a measure of location along a difficulty continuum. The proportion of a group responding correctly to a test item, p, the inverse normal transformation of (1 - p), z, and the item response theory (IRT) difficulty parameter, b, are all used in this way. Using these statistics items can be ordered from very easy to very difficult, and they can be selected to match the characteristics of a particular person or population of persons. However, the ordering of the items on the basis of these statistics may not be the same as the ordering of the items on the basis of difficulty for a particular person. For example, suppose two items with the item characteristic curves shown in Figure 1 are administered to an individual with ability $\theta = .5$. For this person Item A is

---

Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, April 1984.

TM 810 386

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

M. D. Reckase

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

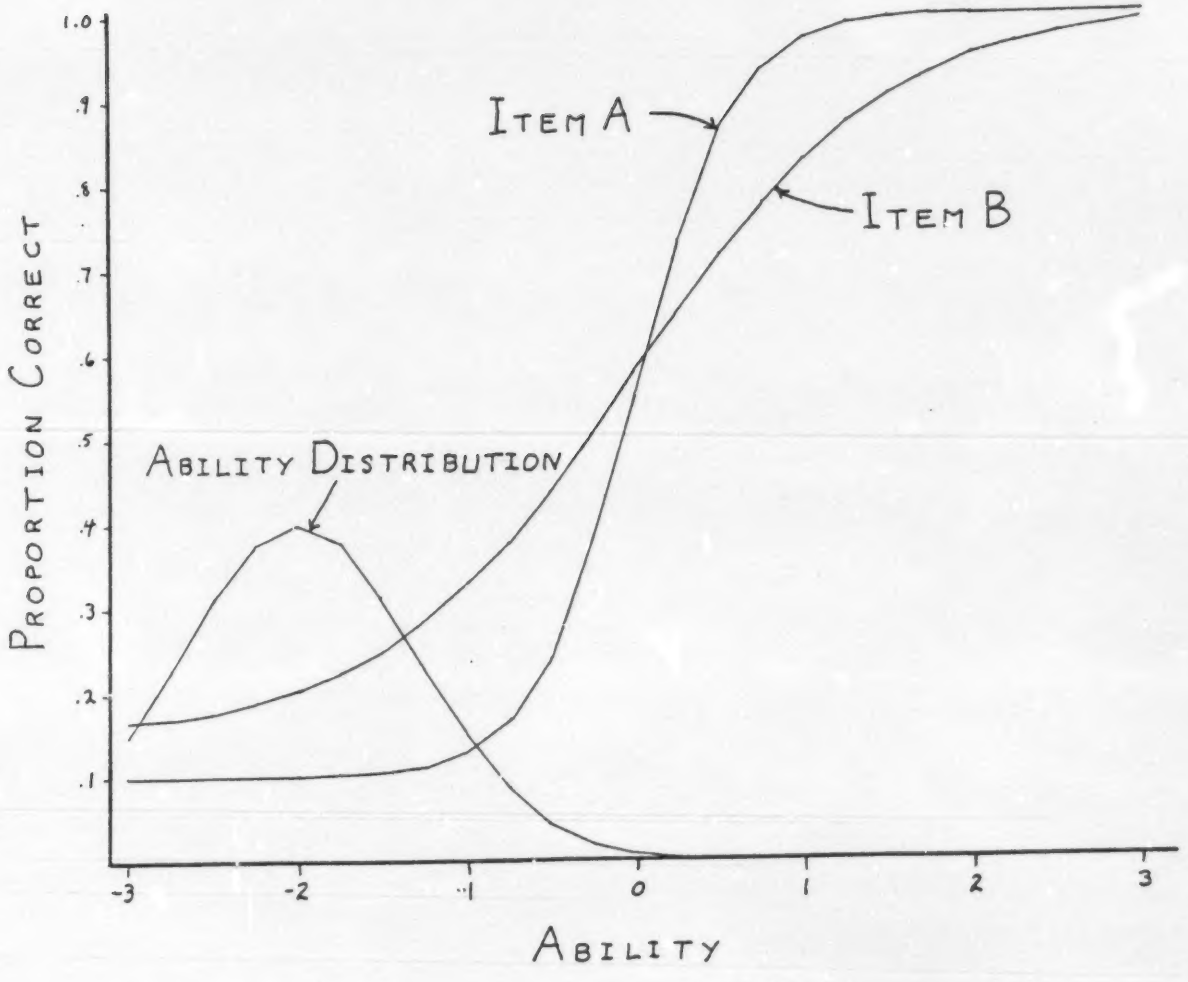
U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

[X] This document has been reproduced as received from the person or organization originating it.
[ ] Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

easier than Item B, but for a group of individuals centered at $\theta = -2$, the p-values of the items are $p_A = .15$ and $p_B = .22$ and the z-values are $z_A = 1.04$ and $z_B = .77$, indicating that Item B is easier than Item A for the group. The b-parameters for the items are both equal to zero. On the basis of the b-parameters, the items are of equal difficulty.

FIGURE 1



These three conceptions of item difficulty have developed independently as a result of different needs. The first is based on a person's perceptions of performance, the second is the traditional, classical test theory approach, and the third is based on the newer IRT methodology. The purpose of this paper is to consider how the indicators of item difficulty are used in practice and to evaluate each of the above procedures to determine how well they meet various test development needs. A new means of indicating item difficulty will then be proposed and evaluated according to the same criteria.

Uses of the Item Difficulty Indexes

Item difficulty indexes have two basic uses. The first is as a description of the difficulty of an item or of a test as a whole. Test publishers commonly publish the average p-value for the items in a test, or even the individual difficulty indices. From this information, an interested practitioner can determine whether the items on the entire test were appropriate for the population on which the statistics were based. He or she can also infer whether the test is of appropriate difficulty for a particular group by comparing the group's performance to the performance of the sample used to compute the item statistics.

The second use of item difficulty indexes is to assist in the selection of items for a test. Items can be selected so that they are neither too hard nor too easy for a target population and the distribution of scores on the test can be controlled by the appropriate selection of test items. In order to use item difficulty statistics for test construction, it must be possible to classify items into various difficulty categories on the basis of the statistics or to rank the items on difficulty.

The commonly used indexes of item difficulty each have some disadvantages when they are used for the above applications. The common p-index of difficulty cannot be easily generalized beyond the population used to compute the values. Also, it is reported on a scale that is quite dissimilar from the scales used to report test scores. This makes it difficult to use this statistic to select an item for a particular ability level. The z-statistic overcomes some of these problems, but it is still not generalizable beyond the population used to compute the values. The probability of a correct response by a particular individual is very difficult to estimate and probably cannot be used for most applications.

Another shortcoming of all of these measures is that they do not tell where an item is effective. The real purpose of using item difficulty statistics for test construction is to select items that will yield information about the abilities of the examinees. The current statistics do not lend themselves to this type of use. Therefore, a new type of difficulty measure is suggested.

A New Indicator of Item Difficulty

The previous analysis of the use of the concept of item difficulty indicated that an index should serve three functions: (1) describe the item; (2) tell where it measures effectively; and (3) classify the item according to difficulty level. One indicator of item difficulty that meets these three criteria is the region of the score scale in which the item is effective.

That is, the region of the scale in which the item gives useful information about an examinee.

This item effectiveness region can be determined fairly easily using item response theory methodology. Suppose, for example, that it is desirable to measure the ability of a group of individuals whose ability distribution has been scaled to be N(0, 1). Further, suppose that the desired standard error of measurement is .3. This means that information at a particular ability level must be approximately equal to $1/.3^2$ or 11.11. If this measurement is to be achieved with about 25 appropriate items for each person, the information acquired from the administration of each item must be at least .44.

If the one-parameter logistic (1PL) model is used to calibrate the items on a test, the region of the score scale that is measured with at least .44 information can be obtained by setting the information function for the model to .44 and solving for the values of θ such that

$$I(\theta) = D^2\Psi[D(\theta - b)] = .44 \ . \qquad (1)$$

When θ is .86 score scale units away from b, the information at that point is equal to .44. Within the range of b - .86 to b + .86, the information is higher than .44. Thus, based on the 1PL model and the desired test characteristics given above, the effectiveness range for an item is $b \pm .86$. To obtain a standard error of measurement at the desired level or less for a person, twenty-five items would have to be selected that had b-values within .86 of a unit of the person's ability.

For the two-parameter logistic model (2PL), the effectiveness range is dependent on the a-parameter for the item. It can be determined from the following formula.

$$(\theta - b) = \frac{\ln\left(\dfrac{a + \sqrt{a^2 - \frac{4I}{D^2}}}{a - \sqrt{a^2 - \frac{4I}{D^2}}}\right)}{Da} \ , \qquad (2)$$

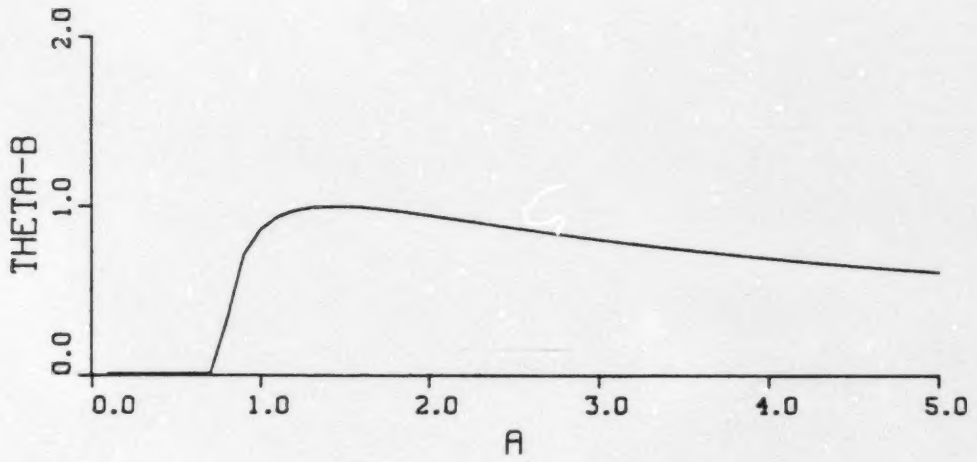
where I is the desired minimum level of information for the item. If an information level of at least .44 is desired for each item, the item must have an a-parameter of at least .78 or Equation 2 yields an imaginary solution. That is, unless $a > .78$ the information provided by the item does not reach .44.

The relationship between $(\theta - b)$ and the a-parameter when an information level of .44 is desired is given in Figure 2. From the graph, it can be seen that in all cases the range of effectiveness of an item is less than $b \pm 1$ and

in most cases is about $\underline{b} \pm .9$. To be safe, a conservative value of $\underline{b} \pm .9$ could be used. In almost every case, the items will have a wider range of effectiveness and will have more information than is required in that range.

Figure 2

Relationship between the a-Parameter
and the Width of the Effectiveness Range
for the Two-Parameter Logistic Model



When guessing is a factor in the item response, the derivation of the effectiveness range of an item becomes more difficult. The relationship between information and ability is given by equation

$$\frac{P^2Q^2}{PQ + cQ^2} = \frac{I}{(1 - c)D^2a^2}, \qquad (3)$$

where $\underline{P}$ is the 2PL expression for the probability of a correct response for a given level of θ, $\underline{Q} = 1 - \underline{P}$, and the other variables have been previously defined. If this equation is solved for $\underline{P}$, $\theta - \underline{b}$ can be determined from

$$\theta - b = \frac{\ln (P/Q)}{D}. \qquad (4)$$

The determination of the values of P that fulfill Equation 3 requires the solution of a cubic equation. While it is not difficult with the use of a computer, it is too complex to be presented here.

For the case when c = .20 and a = 1, and an information of at least .44 is desired, the effectiveness region for an item with guessing present is b - .19 to b + .53. The lack of symmetry is due to the effect of the c-parameter on the shape of the information function. For convenience, a range of b - .20 to b + .50 can be used without any significant loss of precision. Notice that the range of effectiveness for an item with these characteristics is only .70 units, while when guessing is not a factor, the range is 2 x .86 = 1.72. Guessing effects seriously reduce the effectiveness of the items.

The material given above indicates how to determine the range of effectiveness of an item. This information is very useful for selecting test items for a test that is to have particular characteristics. Sometimes it is desirable to have a single measure of the difficulty of the item. In that case the midpoint of the effectiveness range can be used. For items with a c-parameter of zero, the midpoint of the range is simply the b-parameter. For a c-parameter of .2 and an a-parameter of 1, the midpoint is b + .15. This is also very near the point of the θ-scale where the item is maximally informative. In general, the point of maximum information is given by

$$\theta = b + \frac{1}{Da} \ln \left(\frac{1 + \sqrt{1 + 8c}}{2} \right) \qquad (5)$$

(Lord, 1980). This value will be close to the midpoint of the effectiveness range.

The approximate effectiveness range of items using the 1PL, 2PL, and 3PL models is given in Table 1 for the case when a standard error of measurement of .3 is desired for a 25 item test. It is assumed for Table 1 that examinee ability is distributed N(0, 1). These rough approximations to the actual values can be used in test development work.

Table 1

Effectiveness Range and Midpoint
for Items with an a-parameter of 1.0
for the 1PL, 2PL, and 3PL models

Model	Range	Midpoint
1PL	b - .86, b + .86	b
2PL	b - .9, b + .9	b
3PL*	b - .20, b + .50	b + .15

*c-parameter about .20.

Using the Effectiveness Range for Test Construction

Test items can be considered as the building blocks of a test. This is especially true since item information is additive; that is, the information functions for the items in a test are simply summed to get the information function of the test. This fact suggests a test construction procedure that considers that test items are rectangular blocks with length equal to the effectiveness range and height equal to the minimum information provided in the range. For example, a block of information for items calibrated using the 1PL model could be considered to be 1.72 units long and .44 units high. The block is centered at the b-parameter. To construct a test with precision at the desired ability level, "blocks" (items) are simply selected and stacked until the test provides the needed amount of information. If the rough estimates given in this paper are used for this process, the resulting test will generally have more information (greater precision) than is indicated. The information provided in the middle of a "block" is actually much higher than is indicated, since the information function for an item is unimodal rather than rectangular.

An illustration of this test construction process is given below for the case where a pass/fail decision is to be made using a criterion score of -1.0, but reasonable measurement is also desired in the range of ability from -1.0 to + 1.5. At -1.0 a standard error of .3 is required, while a standard error of .5 is sufficient for the rest of the range. The items are assumed to have c-parameters near .2.

Items for the test were first selected to have an effective range midpoint of about $b + .15 = -1$; that is $b = -1.15$. In this way, twenty-five items were selected that had effectiveness ranges that included -1 on the score scale. The information provided by these items results in a standard error of measurement of less than .3 at $\theta = -1$.

In order to obtain a standard error of less than .5 for the remainder of the ability range of interest, at least nine items were required for each ability level. Thus, three sets of nine items each, or 27 items, were selected to cover the region between -1 and 1.5. The items were selected to have no overlap with the initial set of 25 items.

Table 2 shows the item parameters, and effectiveness range for the items selected for the test. Figure 3 shows how these items were stacked to form the test, and Figure 4 shows the resulting information function for the test.

Table 2

Item Parameters and Effectiveness Range

Item	a	b	c	Range	
1	.63	- .97	.11	-1.17	- .47
2	.82	- .81	.11	-1.01	- .31
3	.86	- .85	.11	-1.05	- .35
4	1.23	- .85	.11	-1.05	- .35
5	.70	- .89	.12	-1.09	- .39
6	1.35	- .89	.12	-1.09	- .39
7	.48	-1.34	.12	-1.54	- .84
8	.46	-1.29	.12	-1.49	- .79
9	1.01	- .86	.03	-1.06	- .36
10	.73	- .84	.12	-1.09	- .39
11	.72	-1.09	.12	-1.29	- .59
12	.82	- .87	.12	-1.07	- .37
13	.78	-1.00	.12	-1.20	- .50
14	1.22	- .84	.12	-1.04	- .34
15	.81	-1.02	.12	-1.22	- .52
16	.66	- .86	.12	-1.06	- .36
17	.57	- .98	.12	-1.18	- .48
18	.77	- .68	.12	-1.08	- .38
19	1.07	- .87	.12	-1.07	- .37
20	1.07	-1.19	.11	-1.39	- .69
21	1.09	- .96	.12	-1.16	- .46
22	1.13	-1.21	.12	-1.41	- .71
23	.94	-1.11	.12	-1.31	- .61
24	.82	-1.14	.12	-1.34	- .64
25	.85	-1.06	.12	-1.26	- .56
26	.78	- .25	.12	- .45	.25
27	.96	- .08	.02	- .28	.42
28	.81	- .02	.09	- .22	.48
29	1.08	- .00	.10	- .20	.50
30	1.00	- .21	.19	- .41	.29
31	1.07	- .25	.19	- .45	.25
32	.70	- .62	.12	- .82	- .12
33	1.14	- .64	.11	- .84	- .14
34	.91	- .20	.08	- .40	.30
35	1.14	.46	.22	.26	.96
36	1.09	.60	.12	.46	1.10
37	.53	.71	.09	.51	1.21
38	.93	.67	.25	.47	1.17
39	1.28	.55	.18	.35	1.05
40	1.39	.46	.08	.26	.96
41	1.21	.08	.23	- .12	.58
42	1.01	.06	.08	- .14	.56
43	.89	.52	.16	.32	1.02
44	1.40	1.16	.20	.96	1.66
45	.75	1.43	.35	1.23	1.93
46	1.23	1.47	.26	1.27	1.97
47	2.80	1.47	.14	1.27	1.97
48	.31	1.31	.12	1.11	1.81
49	.72	1.25	.12	1.05	1.75
50	1.56	.86	.20	.66	1.36
51	1.29	.90	.16	.70	1.40
52	1.36	1.24	.21	1.04	1.74

Figure 3

Distribution of Test Items
on Ability Scale

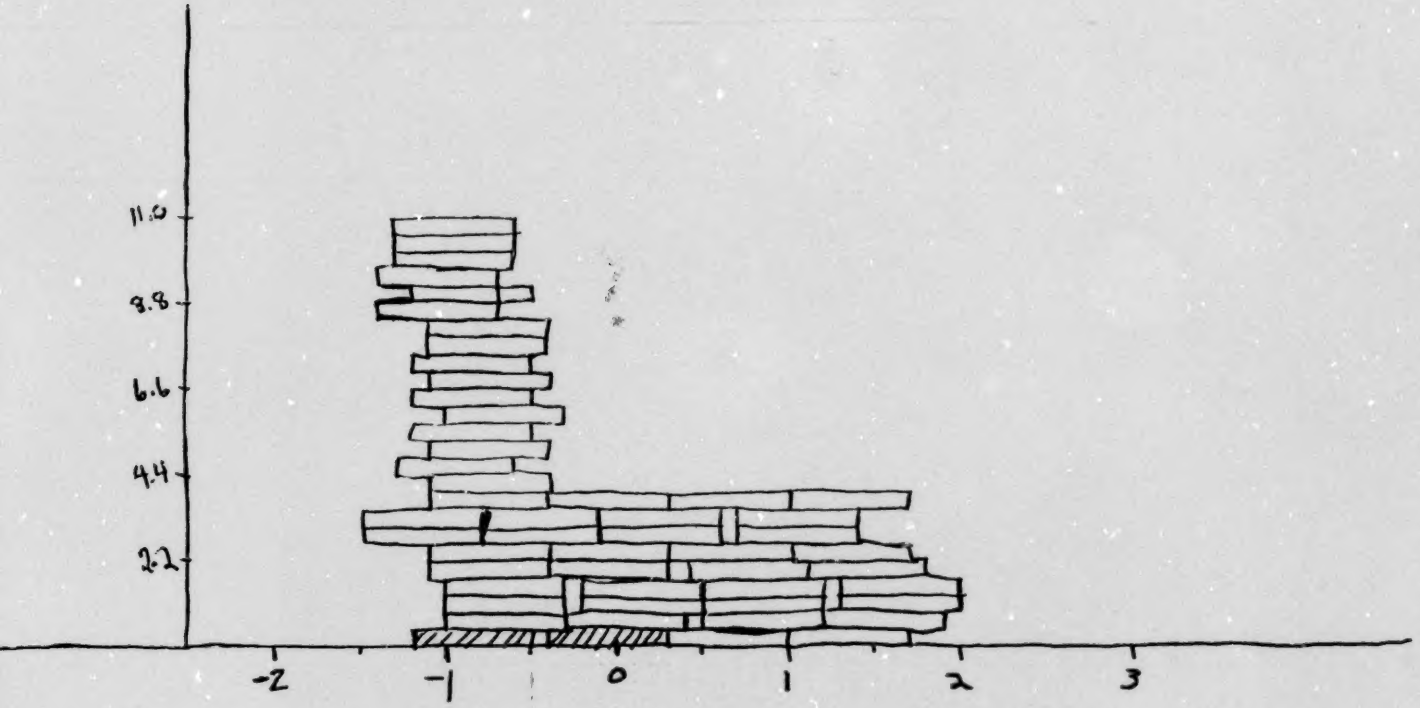
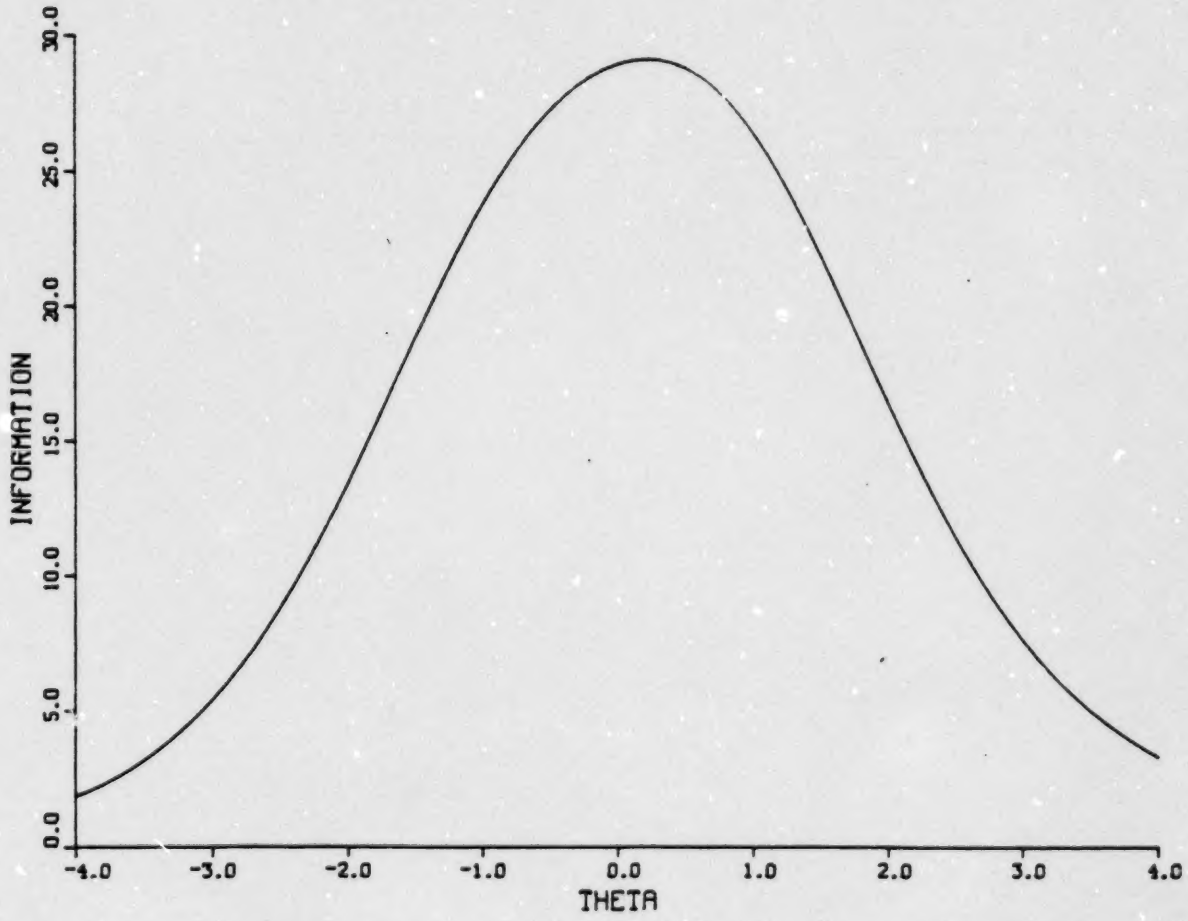


Figure 4

Test Information Function
for Items Presented in Table 2



The test information plot in Figure 4 shows the information at each ability level for the items selected for the test. Notice that at no point is the information less than that specified in Figure 3. The item selection process is conservative in that it underestimates the contribution of each item.

Summary and Conclusions

A method for presenting the difficulty information of an item was presented that included two parts: (a) the range of the score scale in which the item was effective; and (b) the point where the item is most informative. This method of presenting item difficulty is more useful than more traditional means of describing item difficulty because it is sample independent and it can be used more directly for test construction. An example of the use of the effectiveness range for test construction was presented to demonstrate the usefulness of the statistics.

Reference

Lord, F.M. Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum, 1980.

Abstract

Item difficulty reconsidered:
an IRT perspective

The commonly used definition of the concept of item difficulty were summarized and an alternative definition that indicates the range of ability that an item measures effectively was suggested. Formulas for determining the range of effectiveness of an item for three item response theory models were presented along suggested approximate values. An example of using the effectiveness range for test construction was given.